Article

Patterns

Deep learning for detecting and elucidating human T-cell leukemia virus type 1 integration in the human genome

Highlights

- A deep learning model for detecting HTLV-1 VISs was developed
- Bootstrap sampling training strategy improved performance on imbalanced dataset
- Hierarchical clustering of motifs found potential consensus integration sites in humans
- Identified potential *cis*-regulatory features surrounding HTLV-1 VISs

Authors

Haodong Xu, Johnathan Jia, Hyun-Hwan Jeong, Zhongming Zhao

Correspondence

zhongming.zhao@uth.tmc.edu

In brief

Xu and Jia et al. developed a deep learning framework for accurate and rapid detection of human T-cell leukemia virus type 1 (HTLV-1) viral integration sites (VISs) using next-generation sequencing data. Several applications were demonstrated such as motif extraction, discovery of potential integration sites in humans, and identification of transcription factors associated with HTLV-1 integration and disease pathogenesis.

Xu et al., 2023, Patterns *4*, 100674
February 10, 2023
https://doi.org/10.1016/j.patter.2022.100674

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Deep learning for detecting and elucidating human T-cell leukemia virus type 1 integration in the human genome

Haodong Xu,[1,4] Johnathan Jia,[1,2,4] Hyun-Hwan Jeong,[1] and Zhongming Zhao[1,2,3,5,*]
[1]Center for Precision Health, School of Biomedical Informatics, UTHealth Science Center at Houston, Houston, TX 77030, USA
[2]MD Anderson UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA
[4]These authors contributed equally
[5]Lead contact
*Correspondence: zhongming.zhao@uth.tmc.edu
https://doi.org/10.1016/j.patter.2022.100674

THE BIGGER PICTURE Viral integration into the human genome is the cause of several significant diseases such as cancers and latent infections. Accurate detection of viral integration sites (VISs) across the entire genome can be performed rapidly with deep learning. This study presents the first deep learning framework for detecting human T-cell leukemia virus type 1 (HTLV-1) integration sites *de novo* from sequence. Furthermore, we demonstrate how deep learning can provide deeper insight into the *cis*-regulatory features surrounding HTLV-1 VISs. DeepHTLV should be a useful tool for further experimental discovery and validation.

1 2 **3** 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Human T-cell leukemia virus type 1 (HTLV-1), a retrovirus, is the causative agent for adult T cell leukemia/lymphoma and many other human diseases. Accurate and high throughput detection of HTLV-1 virus integration sites (VISs) across the host genomes plays a crucial role in the prevention and treatment of HTLV-1-associated diseases. Here, we developed DeepHTLV, the first deep learning framework for VIS prediction *de novo* from genome sequence, motif discovery, and *cis*-regulatory factor identification. We demonstrated the high accuracy of DeepHTLV with more efficient and interpretive feature representations. Decoding the informative features captured by DeepHTLV resulted in eight representative clusters with consensus motifs for potential HTLV-1 integration. Furthermore, DeepHTLV revealed interesting *cis*-regulatory elements in regulation of VISs that have significant association with the detected motifs. Literature evidence demonstrated nearly half (34) of the predicted transcription factors enriched with VISs were involved in HTLV-1-associated diseases. DeepHTLV is freely available at https://github.com/bsml320/DeepHTLV.

INTRODUCTION

Human T-cell leukemia virus type 1 (HTLV-1), belonging to the genus *Deltaretrovirus*, was the first human retrovirus associated with disease to be identified in the early 1980s. The virus originates from Africa after evolving from zoonosis of simian T lymphotrophic virus.[1] Early infection of HTLV-1 is primarily by cell-to-cell transmission through viral synapses, followed by the virus inserting a DNA copy of its RNA genome into the host cell DNA. Proviral integration sites of HTLV-1 *in vivo* leads to a range of clinical syndromes, and their uncontrolled proliferation is essential for the development of the cancer adult T cell leukemia/lymphoma (ATL), an aggressive $CD4^+$ T cell malignancy.[2] It has been observed that the HTLV-1-encoded viral proteins, e.g., HTLV-1 basic leucine zipper (HBZ) and Tax, play an important role in the development and continued growth of ATL through regulating viral transcription, modulating multiple host transcriptional factors, and perturbing cellular signaling pathways.[3] Approximately 5% of infected individuals develop either ATL- or HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP),[4] a neurodegenerative disorder of the lower limbs. Unfortunately, clinical treatment for HTLV-1 infection and its

associated diseases are lacking and no vaccine currently exists. Therefore, accurate identification of HTLV-1 virus insertion sites (VISs) and their repeatedly inserted genes plays an essential role in the prevention and treatment of diseases.

Until now, diverse experimental methods, e.g., fluorescent *in situ* hybridization and real-time qPCR techniques, have been developed for the detection of VISs, producing substantial useful data for studying VISs.[5,6] Although powerful, none of these molecular biology methods are high throughput and able to detect VISs throughout the whole host genome, whereas HTLV-1-infected hosts can carry anywhere between 500 and 5,000 unique insertion sites with indeterminate preference such as target genes, transcriptional start sites and CpG islands, and transcriptionally silenced regions.[7] In addition, traditional methods are time consuming, resource intensive, and laborious. Therefore, designing and constructing a sensitive and fast VIS detection system is highly needed. To facilitate the rapidly growing number of studies on disease-associated viruses, we have recently developed several computational approaches, e.g., VirusFinder and VERSE, that detect virus integration sites in host genomes based on next-generation sequencing (NGS) of genomic data.[8,9] Moreover, we developed a manually curated VIS database known as the Viral Integration Site Database (VISDB),[10] which contains a large number of virus integration samples taken from experimental research and other resources.[11] Our curated virus integration data provide benchmarks for the development of computational methods to predict potential viral integration sites in the host (human) genome. Tang et al. mined publicly available scientific literature to collect and curate NGS VIS data from several types of studies including experimental studies and other VIS databases such as the Retroviral Integration Database (RID).[11] Specifically, HTLV-1 data from VISDB consisted of experimental identified VISs from four studies. Turpin et al.[12] identified 5,752 VISs experimentally. VISDB data included the VISs found by Cook et al.[13] (11,278 sites), Artesi et al.[14] (4,230 sites), and Furuta et al.[15] (12,585 sites) that were available on RID. We recently released the DeepVISP, a deep learning-based tool that detects multiple oncogenic DNA virus integration (HPV, EBV, and HBV).[16] The prediction performance of DeepVISP is robust: it has area under the curve (AUC) values greater than 0.8 in all the models for all the three DNA viruses. When compared with classical machine learning methods, DeepVISP had an enhancement of 8.43%–34.33% in AUC values. In addition, Hu et al. developed DeepHINT for HIV-1 integration prediction and showed its capability to facilitate the mechanistic studies of the HIV integration process.[17] They could predict HIV VISs with an AUC between 0.736 and 0.904 depending on the dataset. These computational methods have demonstrated that deep learning can be used as an alternative approach to possess sufficient prediction power and provide important biological implication for viral integration prediction.

Here, we developed an interpretable deep neural network (DNN) framework, known as DeepHTLV (Figure 1), for RNA retrovirus integration site prediction *de novo* from genome sequence, motif discovery, and *cis*-regulatory factor identification. Using our curated, largest benchmark integration dataset of 33,845 HTLV-1 VISs, we investigated the insertion tendency regarding chromosome distribution and preferred target genes. We demonstrated the accuracy of DeepHTLV, and its promising performance compared with conventional machine-learning methods, such as decision tree (DT), random forest (RF), K-nearest neighbors (KNN), and logistic regression (LR), by generating more efficient and interpretive feature representations. To improve model performance and avoid false positives, we also implemented a bootstrapping training strategy with 10-fold cross-validation (CV) to generate a multiple ensemble model. The overall performance of the DeepHTLV resulted in AUC values of 0.75- to 10-fold CV. To demonstrate that DeepHTLV is not only accurate but easily interpretable, motifs were extracted from the kernels in the first convolutional layer with the maximum activation. Through clustering the informative motifs captured by DeepHTLV, we discovered eight representative motif clusters with consensus sequences for potential HTLV-1 integration in humans. Furthermore, DeepHTLV revealed interesting *cis*-regulatory patterns around the VISs. Over 70 DNA transcription factor binding sites were found that have significant association with the detected motifs, such as Jun, Fos, and Sp1, while sequence motifs of these TFs were over-represented at the viral insertion sites. Literature evidence demonstrated that 34 of these TFs were involved in either HTLV-1 integration/replication or with HTLV-1 associated diseases, suggesting that DeepHTLV is not only accurate but can make functionally relevant and biologically meaningful predictions. In summary, DeepHTLV is a novel deep learning method that can effectively predict oncogenic retrovirus integration sites and discover the insertion motifs as well as *cis*-regulatory factors, which can be useful for further exploration and understanding of HTLV-1 integration and disease pathogenesis. DeepHTLV is freely available at https://github.com/bsml320/DeepHTLV.

RESULTS

Characterizing HTLV-1 viral integration throughout the genome and training DeepHTLV

A total of 33,845 positive VISs were downloaded from our in-house VISDB.[10] We first investigated the insertion tendency regarding chromosome distribution and preferred target genes (Figures 2A and 2B). We found that HTLV-1 preferentially integrated in the first four chromosomes, with 7.23% in chromosome 1, 7.36% in chromosome 2, 6.93% in chromosome 3, and 7.19% in chromosome 4. Chromosome Y had the smallest number of insertion sites, with 0.27% of all VISs. When normalized by chromosome length, the VISs across the genome were almost uniformly distributed except for chromosome Y. Furthermore, we counted the top 10 genes with the most VISs (Figure 2C). Among them, fragile histidine triad gene (*FHIT*, 23 VISs), a tumor suppressor gene, was a biomarker for early screening of adult T cell leukemia.[18] In addition, for the top 100 genes with the most VISs, we performed gene ontology (GO) and KEGG pathway enrichment analysis by using clusterProfiler[19] (threshold cutoff $p < 0.001$) (Figures 2D and 2E). The results of GO molecular function enrichment indicated that cadherin binding followed by beta-catenin binding (GO:0008013) was the most enriched biological process terms. KEGG pathways were enriched in axon guidance followed by cell adhesion molecules (hsa04514).

Using this largest benchmark integration dataset of VISs, an interpretable deep learning-based predictor, namely DeepHTLV,

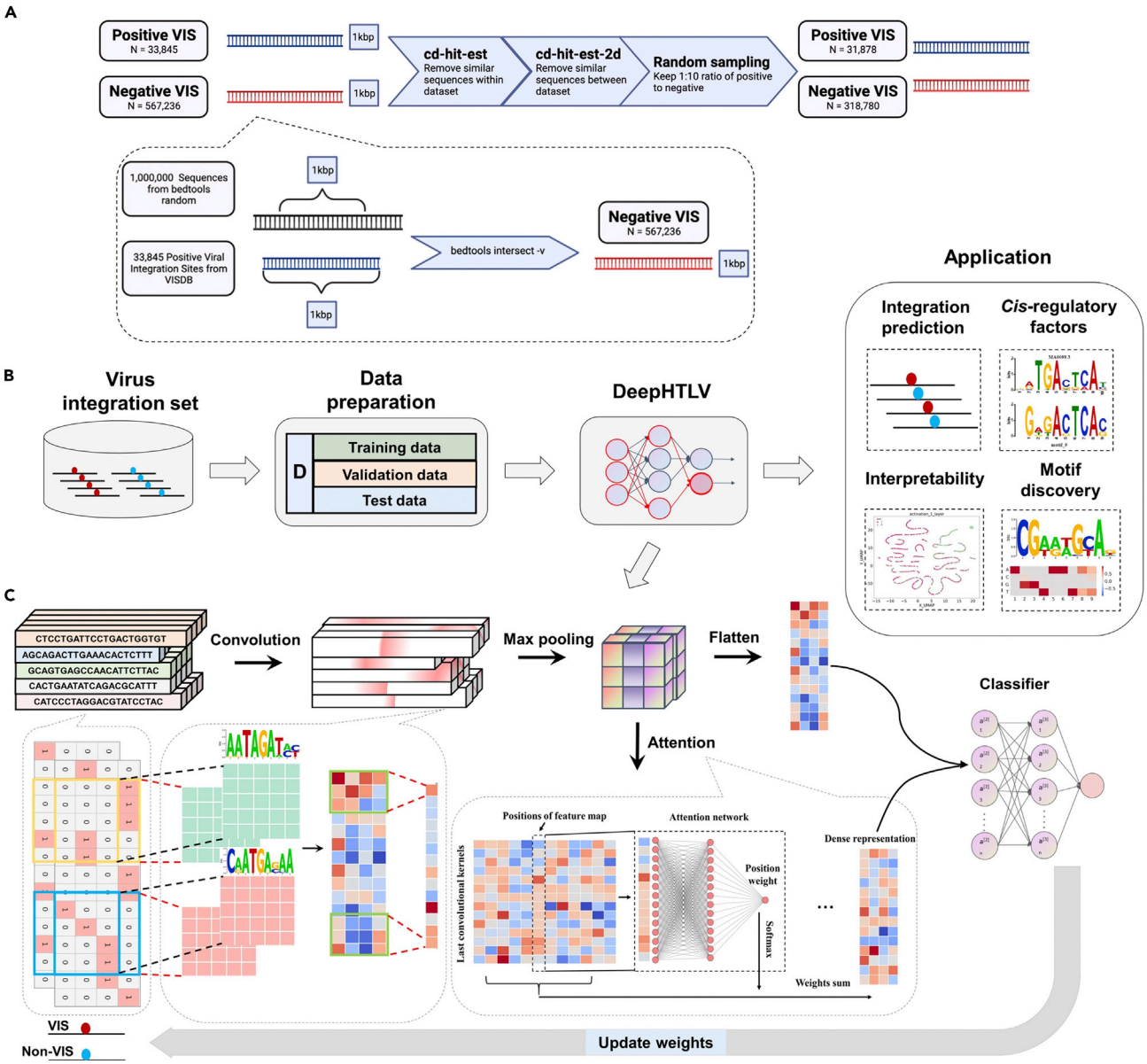

Figure 1. DeepHTLV overview

(A) The dataset retrieved from VISDB was processed with bedtools and CD-HIT to generate 31,878 positive virus integration sites (VISs) and 318,780 non-integration sites for model training and testing.

(B) The whole dataset was split into training and testing datasets (9:1) for DeepHTLV construction and evaluation. DeepHTLV could be used for four different applications: (1) integration site prediction, (2) model interpretability, (3) motif discovery, and (4) *cis*-regulatory element identification.

(C) DeepHTLV was implemented by convolutional neural network (CNN) with attention mechanism. The model input was a matrix consisting of the one-hot-encoded sequence generated after converting the base pairs into binary vectors. Then, the matrix was fed into a convolutional-pooling module, which was followed by an attention architecture. Output from the attention layer was integrated with output from the convolutional-pooling module and sent to a sigmoid activation layer for integration site prediction.

was developed for VIS prediction *de novo* from genome sequence, motif discovery, and *cis*-regulatory factor identification by automatically learning informative features and essential genomic positions (Figures 1B and 1C). Note that the number of non-integration sites was set to be 10 times as many as VISs to mimic the natural imbalance of VISs versus non-integration sites. To improve model performance, we implemented our architecture with a bootstrapping method (Figure S1). All non-integration sites were divided into 10 bins according to the number of VISs. Bootstrap iterations were executed 10 times to generate one classifier. This procedure was repeated to generate 10 classifiers. The average output calculated by all classifiers would be taken as the final prediction. To facilitate the training and evaluation process of our model, the whole dataset was separated into strictly

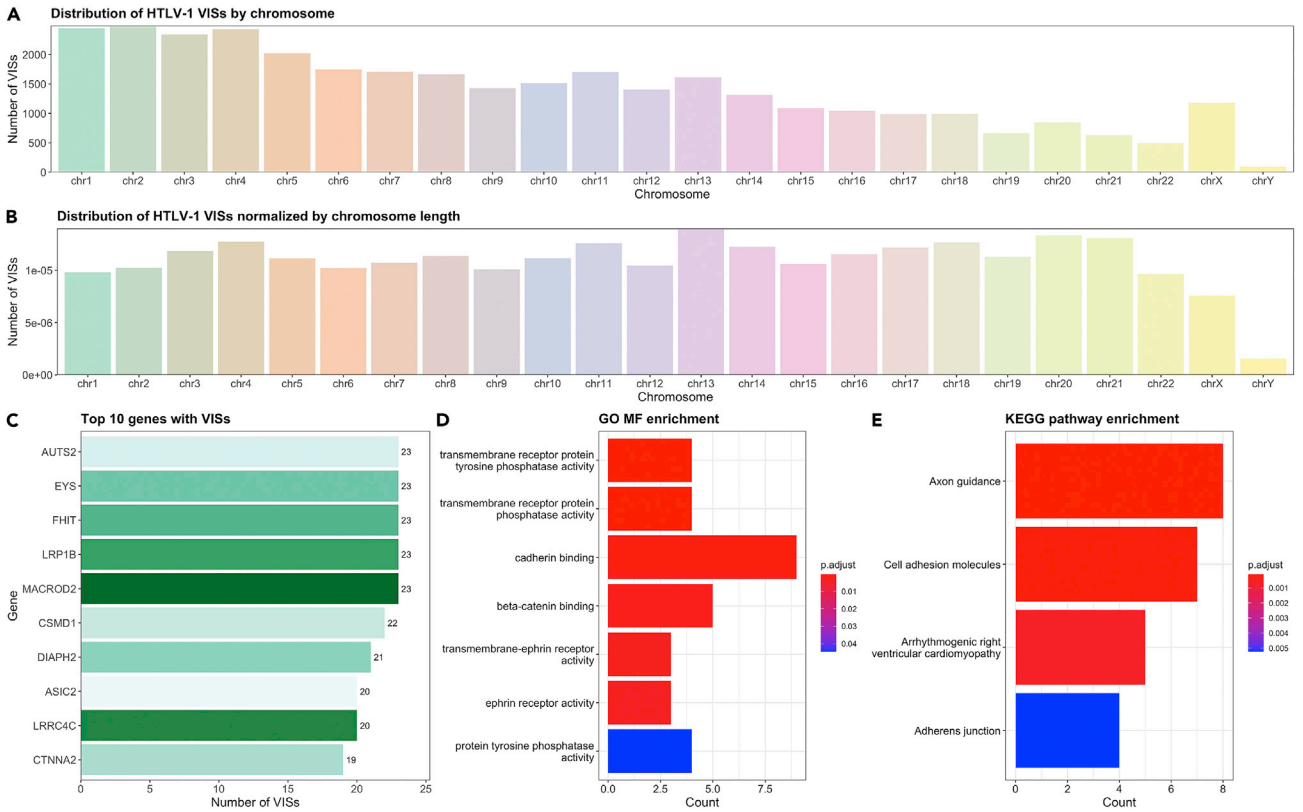

Figure 2. Distribution and features of curated HTLV-1 virus integration sites

(A) The distribution of HTLV-1 VISs across the human chromosomes.

(B) The distribution of HTLV-1 VISs across the human chromosomes after normalization by chromosome length.

(C) Top 10 genes with the most VISs.

(D and E) Gene Ontology (GO) and KEGG pathway enrichment analysis of the top 100 genes with most VISs.

non-overlapping training and testing sets by 10-fold CV. DeepHTLV was implemented with four components, including the input layer, the convolution-maxpooling module, the attention layer, and the output layer. To determine the optimal deep learning model structure, different architectures were evaluated, including a single-layer convolutional neural network (CNN) with an attention mechanism, a two-layer CNN with attention, a single CNN layer without attention, and a three-layer DNN (Figure S2). The models were trained with the same balanced training with 10-fold CV bootstrapping strategy for comparison. Performance measured by AUC indicated that a single CNN layer with attention was the optimal model structure with an average AUC value of 0.75, outperforming other architectures (0.67–0.73).

DeepHTLV accuracy, robustness, and model interpretability

To evaluate the robustness of DeepHTLV, 10-fold CVs on the training dataset were performed and the ROC curves are shown in Figure 3A. DeepHTLV had average AUC values of greater than 0.75, suggesting its good predictive power. Due to the nature of the data imbalance, we evaluated the reliability of predictions from DeepHTLV by checking for true positives. Specifically, we measured DeepHTLV prediction robustness using the area under the precision-recall (AUPR) curve (Figure 3B). DeepHTLV achieved AUPR values ranging from 0.71 to 0.74 during the balanced sample with 10-fold CV training. Moreover, we tested the adaptability of our models using the independent dataset that was not included in training. DeepHTLV obtained an AUC value of 0.75. These good and consistent AUC values between 10-fold CV and independent testing demonstrated the promising accuracy and robustness of DeepHTLV models. Next, we compared the performance of our model with traditional machine learning algorithms. With the same training strategy, four traditional machine learning models were implemented: LR, RF, DT, and KNN (see details in methods; Figure S3). DeepHTLV model performance was superior to all traditional machine learning methods with a modest improvement of 3%–10% (KNN did not show any classification ability) as measured by AUC values on 10-fold CVs.

To determine whether DeepHTLV could be used to extract important VIS features, we sought to assess model interpretability (Figures 3E–3I). We visualized the VISs and non-integration sites using Uniform Manifold Approximation and Projection (UMAP) based on the feature representation prediction as the data went through each layer of the model. More specifically, in the input layer, which is the feature representation of raw data, our result showed no clear separation between the VISs and non-integration sites. However, as it passed through the

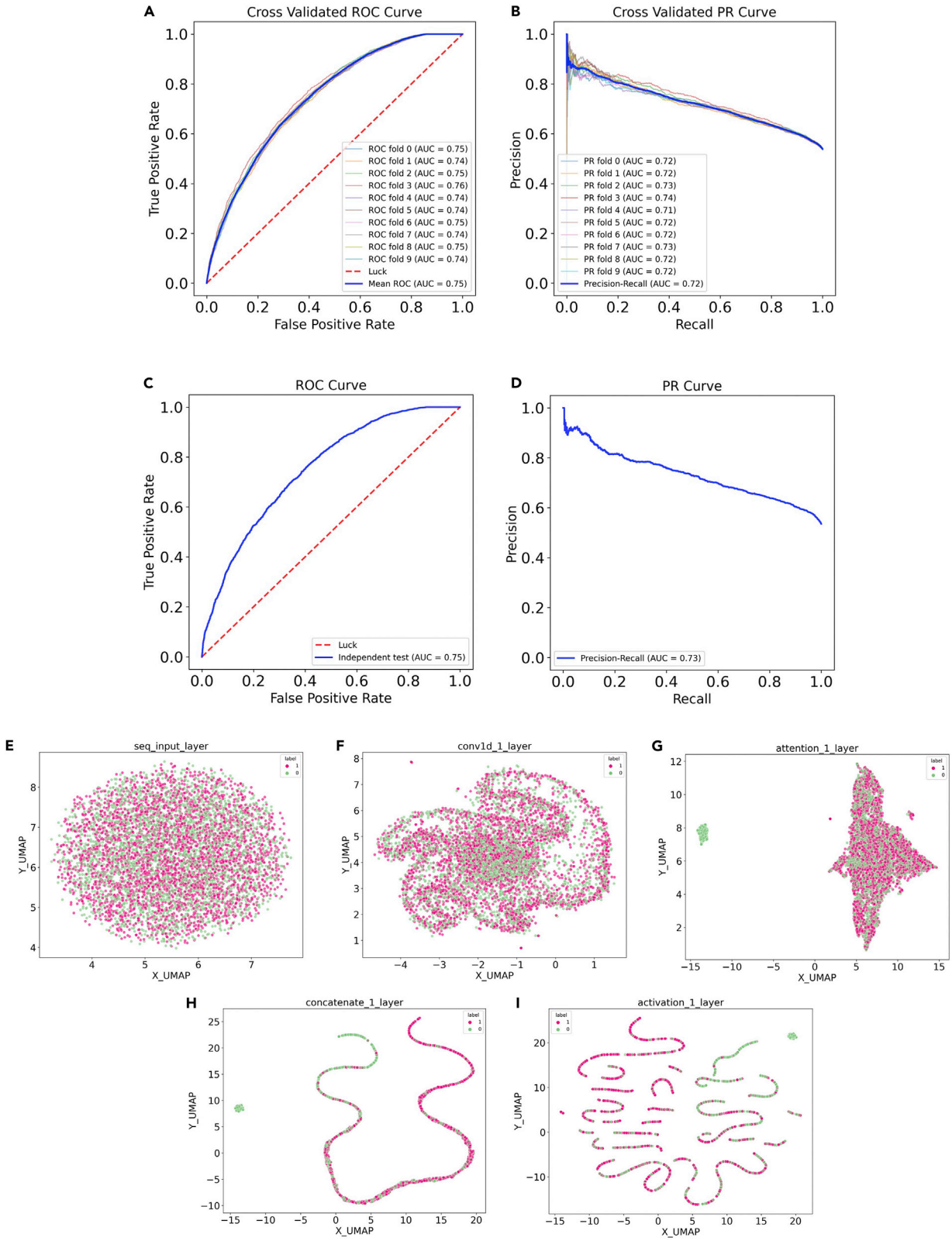

Figure 3. Performance and evaluation for DeepHTLV

(A and B) The area values under the receiver operating characteristic (ROC) curve (AUC) and precision-recall (PR) curve for DeepHTLV were calculated by 10-fold cross-validation (CV) using the bootstrapping strategy.

(legend continued on next page)

CNN layer, the model began differentiating VISs from non-integration sites. The attention layer assigned higher weights to the important genomic positions for determining VISs, and when integrated with the output from the CNN layer the model demonstrated clear separation between VISs and non-integration sites. In the final activation layer, we found that the model retained its ability to distinguish between VISs and non-integration sites. This result indicated that DeepHTLV was capable of learning important genomic features for determining VISs.

DeepHTLV demonstrates consensus motifs potentially important for HTLV-1 integration

The convolutional operation in the CNN is the key operation of the model. Several studies utilized the kernels in the first convolutional layer to extract important sequence motifs.[20] In DeepHTLV, multiple kernels were adopted to determine representative motifs within the input sequences. Each kernel was maximally activated by different regions. By aligning these regions with the input sequences, a position weight matrix (PWM) was generated with the nucleotide count in these corresponding sub-sequences. We considered only the sub-sequences with maximum activation score (MAS) exceeding the threshold (the maximum of the MASs per class). All motifs and the corresponding PWMs were graphically illustrated (Figures S4 and S5). In total, 255 informative motifs were identified. We calculated a score for each motif to measure its importance regarding HTLV-1 integration. For example, the three most important motifs for HTLV-1 integration were "CCCTCTxGA" (Kernel 11, score = 0.12), "CAGTGGTAT" (Kernel 210, score = 0.12), and "AGTAxGTCA" (Kernel 127, score = 0.118). Strong motifs could be detected multiple times because of their importance in HTLV-1 integration. To further explore the dominant patterns of viral insertion, clustering analysis of the top 50 motifs uncovered by DeepHTLV was performed. We first calculated pairwise Spearman's correlations between the PWM of each motif and then applied hierarchical clustering on the correlation matrix, which yields eight representative motif clusters that were potential consensus integration sites (Figure 4). We found that the PWMs in the same cluster tended to have similar core motifs (Figure 4). For example, the core motifs with the consensus sequences CAGTG[GT][AG]T (cluster 6) and x[AC]CTC[CT]x[GC]A (cluster 8) were likely to be involved in the HTLV-1 integration as the top-scoring PWMs; these two motifs were derived from Kernel 210 and Kernel 11, respectively. We further analyzed these two motifs. Histogram plots (Figures 4B and 4C) displayed the maximum activation positions where the motif was extracted, and the violin plots show the distribution of the MASs for VISs and randomly selected non-integration sites. We observed that both motifs had higher average activation scores and larger distributions at the site of viral insertion in the positive samples. Taken together, DeepHTLV was able to detect specific recognition patterns of HTLV-1 insertion sites and revealed consensus motifs potentially important for HTLV-1 integration.

Identifying *cis*-regulatory factors associated with HTLV-1 integration and associated diseases

Physical interactions between viral-encoded multiple proteins, viral transcription factors (TFs), cofactors, host TFs, and other regulators of gene expression are critical steps in viral integration and subsequent replication. These physical interactions generate the necessary machinery that is essential to cause downstream gene expression in the host, which is important to study their roles in human disease pathogenesis. Here, we explored the ability of DeepHTLV to discover and decode some of these interactions by investigating preferential binding of TFs with the motifs that DeepHTLV learned. Among the top 50 learned PWMs, we found 79 TFs whose binding site preference were shown to be significantly associated with the extracted motifs (Figure 5; Table S1) using TOMTOM.[21,22] To verify the accuracy and biological relevance of our predictions, a literature search was conducted to assess the evidence supporting the TF association with HTLV-1 or diseases caused by HTLV-1 infection and integration. Remarkably, 34 TFs[13,23–48] were directly involved with HTLV-1 integration and associated diseases such as ATL and HAM/TSP (Table 1).

Three TFs stood out: Fos, Jun, and specificity protein (Sp). HTLV-1 replication hijacks multiple pathways such as AP-1, NF-κB, and CREB/ATF.[31,49–51] The activation protein 1 (AP-1) signaling pathway regulates several functions including inflammation, cellular proliferation, and apoptosis.[52,53] AP-1 is a dimeric protein complex that involves the recruiting of TFs from the Fos, Jun, ATF, and Maf families. Motif 9 from DeepHTLV predicted multiple TF associations involved in AP-1 signaling from the *FOS* family (*FOS*, *FOSB*, *FOSL1*, and *FOSL2*), the *JUN* family (*JUN*, *JUNB*, and *JUND*), and their respective heterodimer complexes, e.g., *FOSL2-JUND*. Motif 10 extracted from Kernel 160 in DeepHTLV showed significant associations with *MAF*, *MAFK*, and *MAFA*. AP-1 signaling manipulation by HTLV-1 occurred with multiple TFs. Tax and HBZ are two regulatory proteins essential for HTLV-1 replication and disease pathogenesis.[3] Tax interferes with AP-1 signaling on multiple levels. Fujii et al. demonstrated that Tax increased AP-1 activity greater than any combination of the proteins involved in AP-1 signaling.[54] Iwaii et al. reported similar findings: they found that Tax activated genes downstream in the AP-1 pathway by promoting the DNA binding ability the protein complex.[32] Fujii et al. further reported that T cells transformed by HTLV-1 had increased mRNA expression of the AP-1 family members c-Jun, JunB, JunD, c-Fos, and Fra-1, which are the protein products of *JUN*, *JUNB*, *JUND*, *FOS*, and *FOSL1*, respectively. In addition to the predicted Jun and Fos family associations, DeepHTLV predicted Sp1 to have significant associations with motif 17. Sp1 belongs to the family of Kruppel-like/Sp TFs, and is involved in both regulation of gene expression via its ability to bind transcription complexes and chromatin remodeling complexes.[55] Wessner et al. found that Sp1 had binding sites within the U3 region of HTLV-1 LTR.[56] Livengood and Nyborg evaluated the importance of Sp1 HTLV-1 transcription with or without the presence of Tax. Their results indicated that Sp1 could directly bind the viral promoter in multiple regions

(C and D) ROC curves, PR curves, and the AUC values for DeepHTLV using the independent test dataset.
(E–I) Visualization the VISs and non-integration sites using the Uniform Manifold Approximation and Projection (UMAP) based on the feature representation at various network layers. Feature representation of different networks indicated discriminative through the network layer hierarchy.

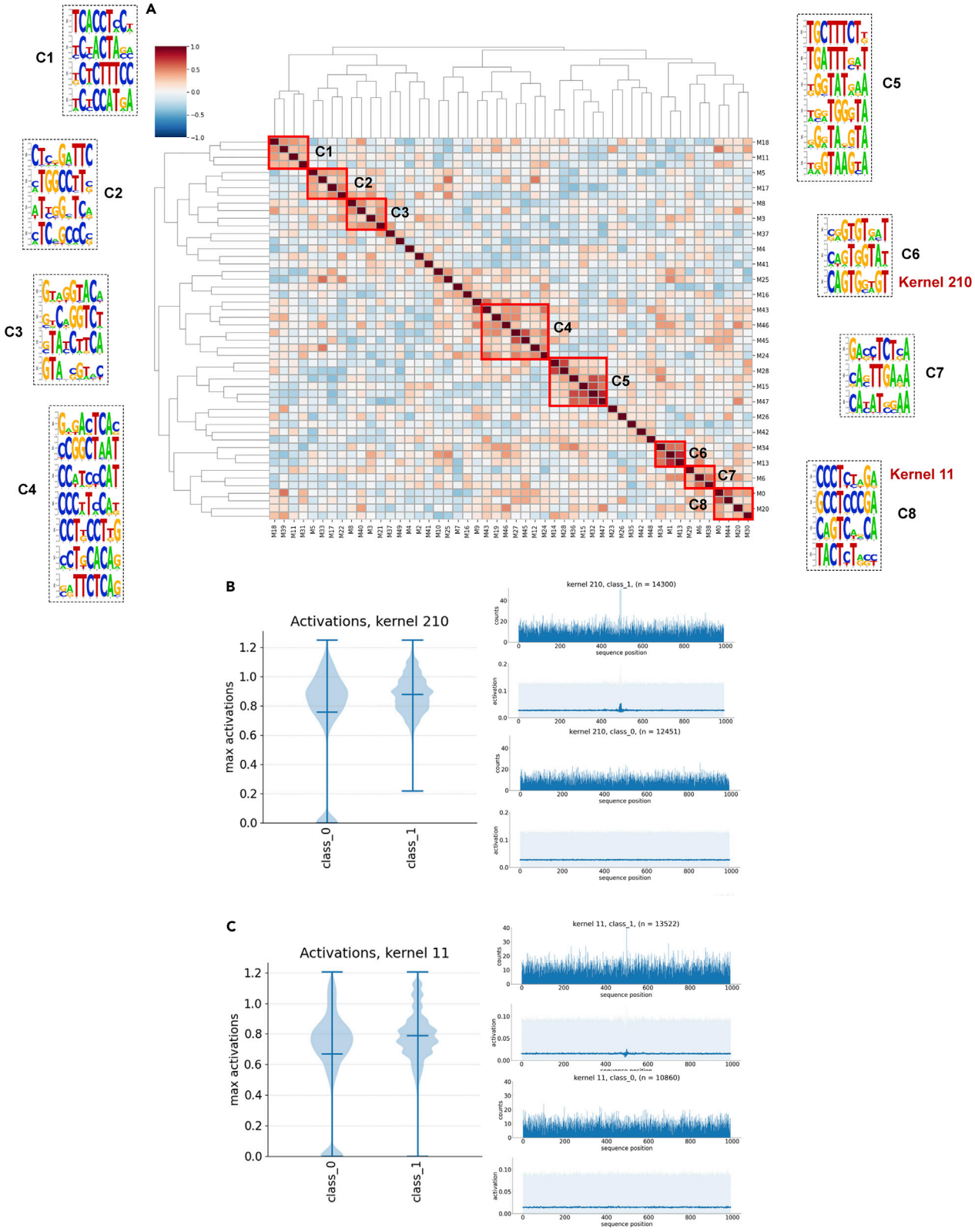

Figure 4. Consensus motifs detected by DeepHTLV

(A) Hierarchical clustering of informative motifs extracted from most activated kernels in the convolutional layer showed eight consensus pattern clusters for potential VISs in humans.

(legend continued on next page)

and that it was most likely involved in Tax-independent basal level transcription.[42] Sp1 in conjunction with Tax converts HTLV-1-infected CD4$^+$ Treg cells into a Th1 profile, promoting IFN-γ production and inflammation in HAM/TSP.[39]

DISCUSSION

In this work, we present the first deep learning model for rapid and accurate HTLV-1 VIS prediction from primary sequence by automatically learning more informative and interpretative features. To improve model accuracy and robustness, we implement a bootstrapping training strategy to enhance the model's ability to discriminate VISs from massive non-integration sites. DeepHTLV outperformed other deep learning architectures and four traditional machine learning models, reaching AUC and AUPR values of 0.75 and 0.72, respectively. So far, some deep learning models have been developed for the VIS prediction of DNA or RNA viruses. For example, DeepHINT, a deep learning method for predicting HIV-1 (another retrovirus) VISs, achieved AUCs between 0.736 and 0.904 depending on the dataset.[17] In addition, using the dataset from the same manually curated database (VISDB), several deep learning-based methods, e.g., DeepVISP, DeepHBV, DeepHPV, and DeepEBV, were used for the VIS prediction of three oncogenic DNA viruses. DeepVISP achieved robust performance with AUC values above 0.8 for all three viruses.[16] Regarding HBV, HPV, and EBV VIS prediction, the performance of DeepHBV, DeepHPV, and DeepEBV with AUC and AUPR values were from 0.610 to 0.794 and 0.547 to 0.574, respectively, using sequence features independently.[57–59] DeepHTLV performance, as indicated by AUPR, outperformed other methods when detecting imbalanced VIS data.

DeepHTLV not only had good performance but also the features it learned were easily interpreted, which was quite important but not available in other methods. Through visualizing the features learned for VISs and non-integration sites using UMAP in each layer of the model, we found that feature representation became more discriminative further along the network layer hierarchy. By decoding these features, DeepHTLV identified several consensus sequence motifs that were important for HTLV-1 integration in humans, such as CAGTG[GT][AG]T and x[AC]CTC[CT]x[GC]A. Furthermore, we demonstrated that DeepHTLV can be used to elucidate the *cis*-regulatory features around HTLV-1 VISs. We found 79 TFs, whose binding site preferences were shown to be significantly associated with the extracted motifs by DeepHTLV. Remarkably, 34 TFs, such as Fos, Jun, and Sp1, were found to have literature evidence supporting their associations with HTLV-1 integration and its associated diseases.

For the 45 remaining predicted TFs, although no literature indicated any relationship with HTLV-1 or diseases caused by viral integration, many of them belong to the same family and/or share sequence homology with one another. For example, Kernel 250, which identified SP1, also identified SP2 and SP4. The relationship between SP1 and HTLV-1 pathogenesis has been previously explored, but the role SP2 and SP4 has not been characterized. All members of the SP family have a highly conserved DNA binding region,[55] and the shared features in this domain could explain why all three were initially identified as associated TFs. More specifically, Sp1 and Sp4 both bind GC boxes while Sp2 only binds GT boxes.[60] The tissue expression between Sp family members differ as well. Sp1 and Sp3 are both expressed in all tissue, while Sp4 is primarily expressed in neuronal cells.[61] Deletion of Sp2 results in the number of neural progenitor cells and neurons decreasing in the cortex, indicating the role of Sp2 in neural development.[62,63] Sp4 is required for proper dendrite formation in the development of the cerebellum.[64] Reduced levels of Sp4 have been observed in bipolar patients.[65] Interestingly, Sp2 was found to be required for *in vitro* expression of RORγT and IL-17 expression in Th17 cells.[66] RORγT is one of the products of the gene *RORC*, which was predicted by DeepHTLV to be a related TF. *RORC*, which encodes for RORγT, showed an age-dependent decrease in expression compared with healthy controls, which suggested a link between IL-17 signaling and ATL.[23] HAM/TSP patients show decreased Th17 count and IL-17 levels compared with asymptomatic carriers and controls, indicating a shift toward a proinflammatory state in HAM/TSP patients.[67] It is possible that Sp2 may play a role in the development of inflammation and subsequent development of disease in HTLV-1-infected patients. Further experimental validation would be required to uncover any possible link. In summary, DeepHTLV was uniquely designed as a deep learning model for retrovirus insertion site prediction and *cis*-regulatory factors identification. We hope this work will contribute to increasing knowledge about viral genomics and for further experimental validation and discovery.

Limitations of the study

Overall, we implemented the simplest possible model, which only used the primary sequence to classify VISs and uncovered their important sequence and motif features. While DeepHTLV made biologically and functionally relevant predictions, there is still room for improvement in performance. Therefore, more features such as structure and expression should be considered to make a more accurate model in the future. In addition, some of the predictions in Table 1 were heterodimers consisting of two different TFs. These heterodimers were reported if at least one of the TFs in the complex had literature support. This assumes that the binding capability of either TF remains unchanged after the heterodimer complex is formed. We are aware that this assumption does not account for conformational changes that may change individual TF binding affinities. However, our model does not consider spatial conformation for the proteins. Rather, it provides a high-confidence reference for potential binding partners of the *cis*-regulatory factors near HTLV-1 VISs. Finally, experimental validation would also be helpful in validating some of the interesting TFs that were predicted by the model that currently lack any supporting literature evidence. How the interplay between VISs, host TFs, and viral proteins affects HTLV-1 disease pathogenesis remains complex and not completely understood.

(B and C) The core motifs: Kernel 210 (B) and Kernel 11 (C), with the consensus sequences CAGTG[GT][AG]T (cluster 6) and x[AC]CTC[CT]x[GC]A (cluster 8). The histograms showed the maximum activation positions where the motif was extracted. The distribution of the maximum activation scores (MASs) for VISs and randomly selected non-integration sites was shown on the violin plots.

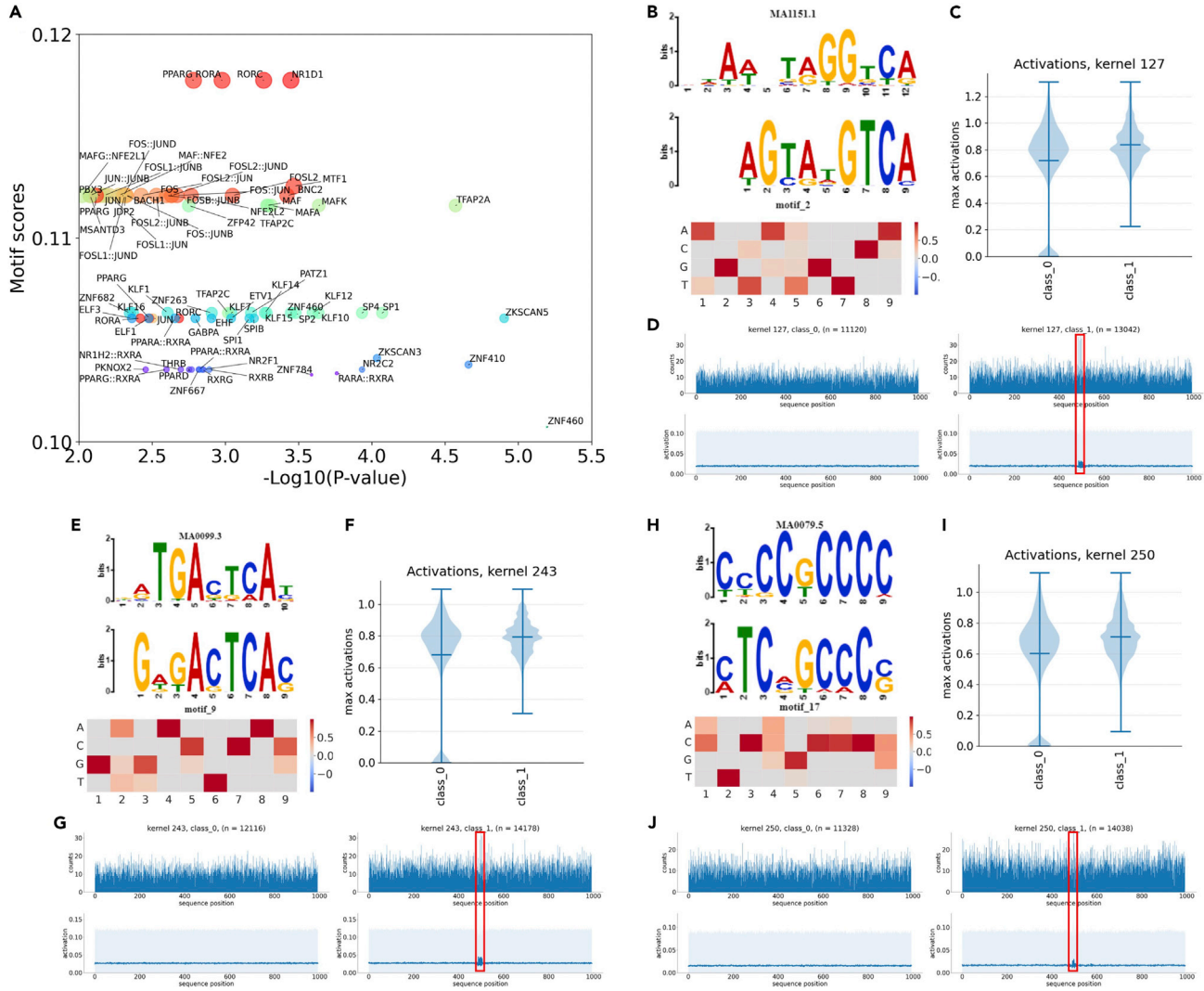

Figure 5. Decoding *cis*-regulatory factors by DeepHTLV

(A) A total of 79 transcription factors (TFs) were identified to match with the top 50 informative motifs extracted from DeepHTLV with statistical significance. $p < 0.01$ was used as the statistical threshold.

(B–J) Graphic sequence, position weight matrix (PWM) and maximum activation distribution for motif 2 (B–D), motif 9 (E–G), and motif 17 (H–J) extracted from DeepHTLV, which matches DNA-binding TF of RORC, FOS-JUN, and SP1 in JASPAR2020 database, respectively. The histograms showed the maximum activation positions where the motif was extracted. The distribution of the MASs for VISs and randomly selected non-integration sites was shown on the violin plots.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Zhongming Zhao (zhongming.zhao@uth.tmc.edu).

Materials availability

No new materials were generated in this study.

Data and code availability

Any requests for additional information are available upon request from the lead contact, Dr. Zhongming Zhao (zhongming.zhao@uth.tmc.edu). DeepHTLV is freely available at https://github.com/bsml320/DeepHTLV. Data used for model training and testing are publicly available on VISDB at https://bioinfo.uth.edu/VISDB/index.php/homepage and available in Table S2.

Data processing

The training dataset of 33,845 experimental VISs was downloaded from our curated VISDB,[10] which was regarded as positive samples (Table S2). For each VIS, we expanded its region by 500 bp upstream and downstream to have a 1,000 bp sequence for feature analysis. Negative samples were generated using bedtools from human genome sequence version GRCh37/hg19, under the constraints that they did not overlap any VISs.[68,69] The ratio of positive and negative samples was set to 1:10 to mimic the natural imbalance of VISs versus non-integration sites, Redundant sequences within each dataset and between both datasets were removed using CD-HIT[70,71] with similarity threshold set to 0.9. After data processing, 31,878 positive samples and 318,780 negative samples remained. The benchmark integration dataset was split into a 9:1 ratio of training to testing data using a package from scikit-learn. Negative samples were randomly selected without replacement at a 1:1 ratio with the positive samples to train models by the bootstrapping strategy. In addition, 727 human TF binding profiles with MEME format[21]

Table 1. Known TFs matched with the top 50 informative motifs learned from DeepHTLV that had literature evidence supporting their associations with HTLV-1 integration and its associated diseases

Transcription factor	Motif/Kernel
RORC[23]	Motif2/Kernel 127, Motif18/Kernel174
PPARG[24]	Motif2/Kernel127, Motif9/Kernel243, Motif18/Kernel174
FOSL2[25,26]	Motif9/Kernel243
FOS-JUN[a,27,28]	Motif9/Kernel243
FOSL2-JUND[a,29]	Motif9/Kernel243
FOS[30,31]	Motif9/Kernel243
FOSB-JUNB[a,30,32]	Motif9/Kernel243
FOSL2-JUNB[a,21,22,26,28]	Motif9/Kernel243
FOSL2-JUN[a,21,22,27]	Motif9/Kernel243
FOS-JUNB[a,30–32]	Motif9/Kernel243
JDP2[33]	Motif9/Kernel243
BACH1[34]	Motif9/Kernel243
FOSL1-JUND[a,25,26,30]	Motif9/Kernel243
FOS-JUND[a,25,26,30,31]	Motif9/Kernel243
JUN[31]	Motif9/Kernel243, Motif18/Kernel174
MAF-NFE2[a,35]	Motif9/Kernel243
FOSL1-JUN[a,30,31]	Motif9/Kernel243
FOSL1-JUNB[a,30,32]	Motif9/Kernel243
JUN-JUNB[a,27,28]	Motif9/Kernel243
MAFG-NFE2L1[a,35]	Motif9/Kernel243
TFAP2A[36]	Motif10/Kernel160
MAF[35]	Motif10/Kernel160
MAFK[37]	Motif10/Kernel160
MAFA[38]	Motif10/Kernel160
SP1[39–42]	Motif17/Kernel250
KLF10[43]	Motif17/Kernel250
KLF12[44]	Motif17/Kernel250
ZNF263[13]	Motif17/Kernel250
SPI1[45]	Motif18/Kernel174
PPARA-RXRA[a,46]	Motif18/Kernel174, Motif29/Kernel223
ELF1[47]	Motif18/Kernel174
NR1H2-RXRA[a,46]	Motif29/Kernel223
PPARG-RXRA[a,20,43]	Motif29/Kernel223
RARA-RXRA[a,48]	Motif31/Kernel213

[a]The motif matching at least one of the TFs in the heterodimer complex.

were downloaded from the JASPAR CORE database (JASPAR 2022) derived from published collections of experimentally defined TF binding sites.[22]

Sequence feature encoding

Sequences were one hot encoded into a binary vector for each nucleotide (ATCG). The binary vectors were then arranged into a matrix with dimensions 4 × N × 1,000, where N is the total number of input samples. A is encoded by (1,0,0,0), C (0,1,0,0), G (0,0,1,0), and T (0,0,0,1). For each VIS, the model input is as follows:

$$BP = (bp_1, bp_2, \ldots, bp_n), bp \in \begin{cases} A(1,0,0,0) \\ C(0,1,0,0) \\ G(0,0,1,0) \\ T(0,0,0,1) \end{cases}$$

$$n \in \{A, C, G, T\}$$

DeepHTLV model construction

As shown in Figure 1C, an attention-based deep learning framework, called DeepHTLV, was developed to predict VISs using nucleotide sequence as input. The architecture of the model consists of eight layers: an input layer, a convolution-pooling module (two layers), a dropout layer, the attention layer, a second dropout layer, a dense layer, and an output layer. The model input is a matrix consisting of the one-hot-encoded sequence data generated after converting the base pairs into binary vectors (sequence feature encoding section). Then, the input matrix is fed into a convolution layer to capture sequence motifs. Each kernel in the convolution operation generates its own PWM, which extracts the important features from the input. For example, for any given VIS region $V(n^1, \ldots, n^{1000})$, the convolution layer ConNet computes:

$$ConNet(V)_{if} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W^f_{mn} V_{i+m,n}$$

where V represents the input sequence, i and f denote the indices of output position and the kernel, respectively. Convolutional kernel W^f is the $M \times N$ weight matrix. M and N are the window size of kernel and input dimension, respectively. For example, N is 4 for the convolutional layer. The activation function is the rectified linear unit (ReLU) being applied to the convolution results, where positive values remain unchanged, and any negative values are set equal to 0. ReLU is defined as:

$$ReLU = \begin{cases} x = x, x \geq 0 \\ x = 0, x \leq 0 \end{cases}$$

where x represents the weight sum of any given neuron. A max-pooling operator was added for dimensional reduction after the convolutional layer. To improve the model performance, an attention layer was used after the max-pooling layer to capture the most valuable sequence motifs. The attention layer takes the features of the convolution-pooling module $\{f_t\}_{t=1}^{T}$ as input and calculates output vector c, suggesting whether the neural network should assign more weights to the positions. c is defined as below:

$$a_t = \frac{\exp(g(f_t))}{\sum_{i=1}^{T} \exp(g(f_i))}$$

$$c = \sum_{t=1}^{T} a_t f_t$$

where g(·) is a neural network with a fully connected layer that returns a scalar importance score. Feature vectors from the convolution-pooling module were merged with the attention scores from the attention layer and fed into the output layer. The output layer is a fully connected dense layer with a sigmoid activation function. The final model output is the probability of whether a given sequence is a VIS.

$$P(x) = \frac{1}{1+e^x}$$

in which x denotes the input of the sigmoid node from the combination of convolution-pooling feature vectors and attention scores.

Hyperparameter optimization was performed using Hyperband from the keras-tuner library.[72] Hyperparameter optimization with Hyperband uses a tournament bracket-style optimization where models are trained briefly, compared, and the more "promising" models are chosen to continue. The model parameters with the best performance determined by AUC value were chosen. In addition, multiple different models were tested, including

models with two convolutional layers plus attention, a single convolutional layer without attention, and a general three-layer DNN. DeepHTLV was implemented using Keras version 2.3.1 and tensorflow-gpu 1.15 to allow for parallel computing in Python version 3.6. When trained on an NVIDIA RTX 3070 with 8 GB of memory, the bootstrap balanced training strategy required 5 h 22 min (322 min) in total. Average training time with 10-fold CV for each ensemble model was approximately 32 or 3.2 min per fold CV. Average DeepHTLV prediction time was 1.3 s.

Model training

All models were trained using a bootstrapping training method (Figure S1) where the negative data were sampled without replacement at a 1:1 ratio with the positive data. After splitting the data into training and testing (9:1 ratio), the negative training data were sampled without replacement with a sample size equal to the positive data (1:1 ratio). These training data were saved as a separate dataset. The model was then trained on the balanced sample with 10-fold CV. In k-fold CV, the training data were separated into k = 10 equally sized partitions where k–1 partitions were used for training and 1 was used for validation. This process was repeated 10 times so that each partition was used for validation once. This process was repeated iteratively until all negative data had been used, resulting in an ensemble of 10 different deep learning models. The final output was determined by taking the average result of all models.

Motif analysis and *cis*-regulatory factors identification

Filters in the convolution layer used a powerful motif detector to scan input sequences as described in DeepBind[73] and Basset.[74] Specifically, sequence motifs were extracted from the convolution layer by finding the positions with the maximum activation. Once the kernels with the maximum activation were determined, they were mapped to the original input sequence to find a set of sub-sequences the length of the filter (kernel size). Only those that exceeded a given maximum activation threshold, i.e., the maximum of the MAS per class, were considered in the subsequent analysis. All sub-sequences were then aligned to generate a PWM to follow the MEME motif format.[21] The motif score was calculated as the difference between the mean maximum activation for positive class and negative class. This score determines how enriched a motif is for the positive class, which in turn corresponds with how important the kernel is for determining whether a sequence is VIS or not. The pysster package was used to generate sequence motif and figures.[75] To further reveal the dominant sequence patterns of viral insertion, clustering analysis of the top 50 motifs DeepHTLV uncovered was performed. The PWMs were grouped by hierarchical clustering based on Spearman's correlations between the nucleic acid composition of PWMs obtained. In addition, after interacting with their host cells, viruses generally dominated the expression of host RNA by virally encoded molecules, which can be realized through physical interactions between a viral transcriptional co-factor and a host TF affecting the downstream host gene expression. Accordingly, we used TOMTOM[21] to compare the motifs that DeepHTLV learned for HTLV-1 integration with the known DNA motifs in JASPAR2022, a database of TF binding profiles.[22] TFs whose binding site preferences were shown to be significantly associated with the extracted motifs were identified. To verify the accuracy and biological relevance of TF predictions, a literature search was performed to determine the evidence supporting the TF association with HTLV-1 or diseases caused by HTLV-1 infection and integration.

Traditional machine learning models

Four traditional machine learning methods were implemented in this study: DT, RF, LR, and KNN. Each model was trained with one-hot-encoded matrix as input. Hyperparameter optimization was performed using RandomizedSearchCV from scikit-learn to find the best classifier from a set of parameters. During hyperparameter optimization, each model went through 10-fold CV. Training data were separated into 10 equally sized partitions where nine parts were used for training and one for evaluation. The ROC curves and PRC were drawn for each model and AUC values were calculated after 10-fold CV to determine the performance for each model. The model performance was calculated as the average of all 10 partitions and used to determine the best parameter set.

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Data		
VIS data	VISDB[10]	https://bioinfo.uth.edu/VISDB/index.php/homepage
Reference Genome	UCSC Genome Browser[69]	https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/
Transcription factor binding profiles	JASPAR 2022[22]	https://jaspar.genereg.net/download/data/2022/CORE/JASPAR2022_CORE_vertebrates_non-redundant_pfms_meme.zip
Software and algorithms		
DeepHTLV	This paper	https://github.com/bsml320/DeepHTLV
Anaconda v4.10.3	Anaconda	https://www.anaconda.com
Python v3.6.13	Python Software Foundation	https://www.python.org
Numpy v1.19.2	Numpy	https://numpy.org
Bedtools v2.30.0	Bedtools	https://github.com/arq5x/bedtools2
Matplotlib v3.3.4	Matplotlib	https://matplotlib.org
Keras v2.3.1 (GPU)	Keras	https://keras.io/
Tensorflow-gpu v1.15.0	Tensorflow	https://www.tensorflow.org
Scikit-learn v0.24.2	Scikit-Learn	https://scikit-learn.org/stable/
CUDA v10.0	NVIDIA	https://developer.nvidia.com/cuda-toolkit
clusterProfiler v 4.0	clusterProfiler[19]	https://bioconductor.org/packages/release/bioc/vignettes/clusterProfiler/inst/doc/clusterProfiler.html
pysster	pysster[75]	https://github.com/budach/pysster
CD-HIT	CD-HIT[70,71]	https://github.com/weizhongli/cdhit/wiki
Keras Tuner	Keras-tuner[72]	https://github.com/keras-team/keras-tuner

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100674.

ACKNOWLEDGMENTS

We would like to thank Dr. Deyou Tang for his help with accessing and using the data from VISDB. Figure 1 was made using BioRender. Z.Z. was partially

supported by the National Institutes of Health grants (R01LM012806, R01DE030122, and R03AG077191). The authors thank for the technical support from the Cancer Genomics Core funded by the Cancer Prevention and Research Institute of Texas (CPRIT RP180734) and the CPRIT BIG-TCR program (RP210045). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Z.Z. conceptualized the project. H.X. and J.J. designed the study and computational framework and analyzed the data and results. H.-H.J. contributed the prototype of the model and participated in some of the data analysis. H.X. and J.J. made the figures and tables and implemented the model. H.X. and J.J. wrote and edited the manuscript. Z.Z. edited the manuscript. Z.Z. helped plan and supervise the research. All authors proofread and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 1, 2022
Revised: November 2, 2022
Accepted: December 13, 2022
Published: February 10, 2023

REFERENCES

1. Mahieux, R., Ibrahim, F., Mauclere, P., Herve, V., Michel, P., Tekaia, F., Chappey, C., Garin, B., Van Der Ryst, E., Guillemain, B., et al. (1997). Molecular epidemiology of 58 new African human T-cell leukemia virus type 1 (HTLV-1) strains: identification of a new and distinct HTLV-1 molecular subtype in Central Africa and in Pygmies. J. Virol. *71*, 1317–1333. https://doi.org/10.1128/jvi.71.2.1317-1333.1997.

2. Kamihira, S., Sugahara, K., Tsuruda, K., Minami, S., Uemura, A., Akamatsu, N., Nagai, H., Murata, K., Hasegawa, H., Hirakata, Y., et al. (2005). Proviral status of HTLV-1 integrated into the host genomic DNA of adult T-cell leukemia cells. Clin. Lab. Haematol. *27*, 235–241. https://doi.org/10.1111/j.1365-2257.2005.00698.x.

3. Doi, K., Wu, X., Taniguchi, Y., Yasunaga, J.-I., Satou, Y., Okayama, A., Nosaka, K., and Matsuoka, M. (2005). Preferential selection of human T-cell leukemia virus type I provirus integration sites in leukemic versus carrier states. Blood *106*, 1048–1053. https://doi.org/10.1182/blood-2004-11-4350.

4. Meekings, K.N., Leipzig, J., Bushman, F.D., Taylor, G.P., and Bangham, C.R.M. (2008). HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. PLoS Pathog. *4*, e1000027. https://doi.org/10.1371/journal.ppat.1000027.

5. Schmidt, M., Schwarzwaelder, K., Bartholomae, C.C., Glimm, H., and von Kalle, C. (2009). Detection of retroviral integration sites by linear amplification-mediated PCR and tracking of individual integration clones in different samples. Methods Mol. Biol. *506*, 363–372. https://doi.org/10.1007/978-1-59745-409-4_24.

6. Redmond, C.J., Fu, H., Aladjem, M.I., and McBride, A.A. (2018). Human papillomavirus integration: analysis by molecular combing and fiber-FISH. Curr. Protoc. Microbiol. *51*, e61. https://doi.org/10.1002/cpmc.61.

7. Gillet, N.A., Malani, N., Melamed, A., Gormley, N., Carter, R., Bentley, D., Berry, C., Bushman, F.D., Taylor, G.P., and Bangham, C.R.M. (2011). The host genomic environment of the provirus determines the abundance of HTLV-1–infected T-cell clones. Blood *117*, 3113–3122. https://doi.org/10.1182/blood-2010-10-312926.

8. Wang, Q., Jia, P., and Zhao, Z. (2013). VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS One *8*, e64465. https://doi.org/10.1371/journal.pone.0064465.

9. Wang, Q., Jia, P., and Zhao, Z. (2015). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. *7*, 2. https://doi.org/10.1186/s13073-015-0126-6.

10. Tang, D., Li, B., Xu, T., Hu, R., Tan, D., Song, X., Jia, P., and Zhao, Z. (2020). VISDB: a manually curated database of viral integration sites in the human genome. Nucleic Acids Res. *48*, D633–D641. https://doi.org/10.1093/nar/gkz867.

11. Shao, W., Shan, J., Kearney, M.F., Wu, X., Maldarelli, F., Mellors, J.W., Luke, B., Coffin, J.M., and Hughes, S.H. (2016). Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. Retrovirology *13*, 47. https://doi.org/10.1186/s12977-016-0277-6.

12. Turpin, J., Yurick, D., Khoury, G., Pham, H., Locarnini, S., Melamed, A., Witkover, A., Wilson, K., Purcell, D., Bangham, C.R.M., and Einsiedel, L. (2019). Impact of hepatitis B virus coinfection on human T-lymphotropic virus type 1 clonality in an indigenous population of Central Australia. J. Infect. Dis. *219*, 562–567. https://doi.org/10.1093/infdis/jiy546.

13. Cook, L.B., Melamed, A., Niederer, H., Valganon, M., Laydon, D., Foroni, L., Taylor, G.P., Matsuoka, M., and Bangham, C.R.M. (2014). The role of HTLV-1 clonality, proviral structure, and genomic integration site in adult T-cell leukemia/lymphoma. Blood *123*, 3925–3931. https://doi.org/10.1182/blood-2014-02-553602.

14. Artesi, M., Marçais, A., Durkin, K., Rosewick, N., Hahaut, V., Suarez, F., Trinquand, A., Lhermitte, L., Asnafi, V., Avettand-Fenoel, V., et al. (2017). Monitoring molecular response in adult T-cell leukemia by high-throughput sequencing analysis of HTLV-1 clonality. Leukemia *31*, 2532–2535. https://doi.org/10.1038/leu.2017.260.

15. Furuta, R., Yasunaga, J.I., Miura, M., Sugata, K., Saito, A., Akari, H., Ueno, T., Takenouchi, N., Fujisawa, J.I., Koh, K.R., et al. (2017). Human T-cell leukemia virus type 1 infects multiple lineage hematopoietic cells in vivo. PLoS Pathog. *13*, e1006722. https://doi.org/10.1371/journal.ppat.1006722.

16. Xu, H., Jia, P., and Zhao, Z. (2021). DeepVISP: deep learning for virus site integration prediction and motif discovery. Adv. Sci. *8*, 2004958. https://doi.org/10.1002/advs.202004958.

17. Hu, H., Xiao, A., Zhang, S., Li, Y., Shi, X., Jiang, T., Zhang, L., Zhang, L., and Zeng, J. (2019). DeepHINT: understanding HIV-1 integration via deep learning with attention. Bioinformatics *35*, 1660–1667. https://doi.org/10.1093/bioinformatics/bty842.

18. Bellon, M., Bialuk, I., Galli, V., Bai, X.-T., Farre, L., Bittencourt, A., Marçais, A., Petrus, M.N., Ratner, L., Waldmann, T.A., et al. (2021). Germinal epimutation of Fragile Histidine Triad (FHIT) gene is associated with progression to acute and chronic adult T-cell leukemia diseases. Mol. Cancer *20*, 86. https://doi.org/10.1186/s12943-021-01370-2.

19. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

20. Chen, S., Gan, M., Lv, H., and Jiang, R. (2021). DeepCAPE: a deep convolutional neural network for the accurate prediction of enhancers. Dev. Reprod. Biol. *19*, 565–577. https://doi.org/10.1016/j.gpb.2019.04.006.

21. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). Meme SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208. https://doi.org/10.1093/nar/gkp335.

22. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). Jaspar 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *48*, D87–D92. https://doi.org/10.1093/nar/gkz1001.

23. Subramanian, K., Dierckx, T., Khouri, R., Menezes, S.M., Kagdi, H., Taylor, G.P., Farre, L., Bittencourt, A., Kataoka, K., Ogawa, S., and Van Weyenbergh, J. (2019). Decreased RORC expression and downstream signaling in HTLV-1-associated adult T-cell lymphoma/leukemia uncovers an antiproliferative IL17 link: a potential target for immunotherapy? Int. J. Cancer *144*, 1664–1675. https://doi.org/10.1002/ijc.31922.

24. Zarei-Ghobadi, M., Sheikhi, M., Teymoori-Rad, M., Yaslianifard, S., Norouzi, M., Yaslianifard, S., Faraji, R., Farahmand, M., Bayat, S., Jafari, M., and Mozhgani, S.-H. (2021). HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) versus adult T-cell leukemia/lymphoma (ATLL). BMC Res. Notes *14*, 109. https://doi.org/10.1186/s13104-021-05521-y.

25. Nakayama, T., Hieshima, K., Arao, T., Jin, Z., Nagakubo, D., Shirakawa, A.K., Yamada, Y., Fujii, M., Oiso, N., Kawada, A., et al. (2008). Aberrant expression of Fra-2 promotes CCR4 expression and cell proliferation in adult T-cell leukemia. Oncogene *27*, 3221–3232. https://doi.org/10.1038/sj.onc.1210984.

26. Terol, M., Gazon, H., Lemasson, I., Duc-Dodon, M., Barbeau, B., Césaire, R., Mesnard, J.M., and Péloponèse, J.M., Jr. (2017). HBZ-mediated shift of JunD from growth suppressor to tumor promoter in leukemic cells by inhibition of ribosomal protein S25 expression. Leukemia *31*, 2235–2243. https://doi.org/10.1038/leu.2017.74.

27. Fu, W., Shah, S.R., Jiang, H., Hilt, D.C., Dave, H.P., and Joshi, J.B. (1997). Transactivation of proenkephalin gene by HTLV-1 tax1 protein in glial cells: involvement of Fos/Jun complex at an AP-1 element in the proenkephalin gene promoter. J. Neurovirol. *3*, 16–27. https://doi.org/10.3109/13550289709015789.

28. Jeang, K.-T., Chiu, R., Santos, E., and Kim, S.-J. (1991). Induction of the HTLV-I LTR by Jun occurs through the tax-responsive 21-bp elements. Virology *181*, 218–227. https://doi.org/10.1016/0042-6822(91)90487-V.

29. Nakayama, T., Higuchi, T., Oiso, N., Kawada, A., and Yoshie, O. (2012). Expression and function of FRA2/JUND in cutaneous T-cell lymphomas. Anticancer Res. *32*, 1367–1373.

30. Fujii, M., Niki, T., Mori, T., Matsuda, T., Matsui, M., Nomura, N., and Seiki, M. (1991). HTLV-1 Tax induces expression of various immediate early serum responsive genes. Oncogene *6*, 1023–1029.

31. Gazon, H., Barbeau, B., Mesnard, J.-M., and Peloponese, J.-M. (2017). Hijacking of the AP-1 signaling pathway during development of ATL. Front. Microbiol. *8*, 2686. https://doi.org/10.3389/fmicb.2017.02686.

32. Iwai, K., Mori, N., Oie, M., Yamamoto, N., and Fujii, M. (2001). Human T-cell leukemia virus type 1 tax protein activates transcription through AP-1 site by inducing DNA binding activity in T cells. Virology *279*, 38–46. https://doi.org/10.1006/viro.2000.0669.

33. Nakano, K., Yokoyama, K., Shin, S., Uchida, K., Tsuji, K., Tanaka, M., Uchimaru, K., and Watanabe, T. (2022). Exploring new functional aspects of HTLV-1 RNA-binding protein Rex: how does Rex control viral replication? Viruses *14*, 407.

34. Fochi, S., Ciminale, V., Trabetti, E., Bertazzoni, U., D'Agostino, D.M., Zipeto, D., and Romanelli, M.G. (2019). NF-κB and MicroRNA deregulation mediated by HTLV-1 tax and HBZ. Pathogens *8*, 290.

35. Reinke, A.W., Grigoryan, G., and Keating, A.E. (2010). Identification of bZIP interaction partners of viral proteins HBZ, MEQ, BZLF1, and K-bZIP using coiled-coil arrays. Biochemistry *49*, 1985–1997. https://doi.org/10.1021/bi902065k.

36. Sasaki, H., Nishikata, I., Shiraga, T., Akamatsu, E., Fukami, T., Hidaka, T., Kubuki, Y., Okayama, A., Hamada, K., Okabe, H., et al. (2005). Overexpression of a cell adhesion molecule, TSLC1, as a possible molecular marker for acute-type adult T-cell leukemia. Blood *105*, 1204–1213. https://doi.org/10.1182/blood-2004-03-1222.

37. Rushing, A.W., Rushing, B., Hoang, K., Sanders, S.V., Péloponèse, J.M., Jr., Polakowski, N., and Lemasson, I. (2019). HTLV-1 basic leucine zipper factor protects cells from oxidative stress by upregulating expression of Heme Oxygenase I. PLoS Pathog. *15*, e1007922. https://doi.org/10.1371/journal.ppat.1007922.

38. Pinto, M.T., Malta, T.M., Rodrigues, E.S., Pinheiro, D.G., Panepucci, R.A., Malmegrim de Farias, K.C.R., Sousa, A.D.P., Takayanagui, O.M., Tanaka, Y., Covas, D.T., and Kashima, S. (2014). Genes related to antiviral activity, cell migration, and lysis are differentially expressed in CD4(+) T cells in human t cell leukemia virus type 1-associated myelopathy/tropical spastic paraparesis patients. AIDS Res. Hum. Retrovir. *30*, 610–622. https://doi.org/10.1089/aid.2013.0109.

39. Araya, N., Sato, T., Ando, H., Tomaru, U., Yoshida, M., Coler-Reilly, A., Yagishita, N., Yamauchi, J., Hasegawa, A., Kannagi, M., et al. (2014). HTLV-1 induces a Th1-like state in CD4+CCR4+ T cells. J. Clin. Invest. *124*, 3431–3442. https://doi.org/10.1172/JCI75250.

40. Barbeau, B., Peloponese, J.-M., and Mesnard, J.-M. (2013). Functional comparison of antisense proteins of HTLV-1 and HTLV-2 in viral pathogenesis. Front. Microbiol. *4*, 226. https://doi.org/10.3389/fmicb.2013.00226.

41. Gazon, H., Lemasson, I., Polakowski, N., Césaire, R., Matsuoka, M., Barbeau, B., Mesnard, J.M., and Peloponese, J.M., Jr. (2012). Human T-cell leukemia virus type 1 (HTLV-1) bZIP factor requires cellular transcription factor JunD to upregulate HTLV-1 antisense transcription from the 3' long terminal repeat. J. Virol. *86*, 9070–9078. https://doi.org/10.1128/jvi.00661-12.

42. Livengood, J.A., and Nyborg, J.K. (2004). The high-affinity Sp1 binding site in the HTLV-1 promoter contributes to Tax-independent basal expression. Nucleic Acids Res. *32*, 2829–2837. https://doi.org/10.1093/nar/gkh590.

43. Taniguchi, H., Hasegawa, H., Sasaki, D., Ando, K., Sawayama, Y., Imanishi, D., Taguchi, J., Imaizumi, Y., Hata, T., Tsukasaki, K., et al. (2014). Heat shock protein 90 inhibitor NVP-AUY922 exerts potent activity against adult T-cell leukemia–lymphoma cells. Cancer Sci. *105*, 1601–1608. https://doi.org/10.1111/cas.12540.

44. Rosewick, N., Durkin, K., Artesi, M., Marçais, A., Hahaut, V., Griebel, P., Arsic, N., Avettand-Fenoel, V., Burny, A., Charlier, C., et al. (2017). Cis-perturbation of cancer drivers by the HTLV-1/BLV proviruses is an early determinant of leukemogenesis. Nat. Commun. *8*, 15264. https://doi.org/10.1038/ncomms15264.

45. Tsukada, J., Misago, M., Serino, Y., Ogawa, R., Murakami, S., Nakanishi, M., Tonai, S., Kominato, Y., Morimoto, I., Auron, P.E., and Eto, S. (1997). Human T-cell leukemia virus type I tax transactivates the promoter of human prointerleukin-1β gene through association with two transcription factors, nuclear factor–interleukin-6 and Spi-1. Blood *90*, 3142–3153. https://doi.org/10.1182/blood.V90.8.3142.

46. Lin, J.H., Kim, E.J., Bansal, A., Seykora, J., Richardson, S.K., Cha, X.-Y., Zafar, S., Nasta, S., Wysocka, M., Benoit, B., et al. (2008). Clinical and in vitro resistance to bexarotene in adult T-cell leukemia: loss of RXR-α receptor. Blood *112*, 2484–2488. https://doi.org/10.1182/blood-2008-03-141424.

47. Clark, N.M., Smith, M.J., Hilfinger, J.M., and Markovitz, D.M. (1993). Activation of the human T-cell leukemia virus type I enhancer is mediated by binding sites for Elf-1 and the pets factor. J. Virol. *67*, 5522–5528. https://doi.org/10.1128/jvi.67.9.5522-5528.1993.

48. Fujimura, S., Suzumiya, J., Anzai, K., Ohkubo, K., Hata, T., Yamada, Y., Kamihira, S., Kikuchi, M., and Ono, J. (1998). Retinoic acids induce growth inhibition and apoptosis in adult T-cell leukemia (ATL) cell lines. Leuk. Res. *22*, 611–618. https://doi.org/10.1016/S0145-2126(98)00049-6.

49. Martin, J.L., Maldonado, J.O., Mueller, J.D., Zhang, W., and Mansky, L.M. (2016). Molecular studies of HTLV-1 replication: an update. Viruses *8*, 31. https://doi.org/10.3390/v8020031.

50. Matsuoka, M., and Jeang, K.T. (2011). Human T-cell leukemia virus type 1 (HTLV-1) and leukemic transformation: viral infectivity, Tax, HBZ and therapy. Oncogene *30*, 1379–1389. https://doi.org/10.1038/onc.2010.537.

51. Yao, J., and Wigdahl, B. (2000). Human T cell lymphotropic virus type I genomic expression and impact on intracellular signaling pathways during neurodegenerative disease and leukemia. Front. Biosci. *5*, D138–D168. https://doi.org/10.2741/yao.

52. Eferl, R., and Wagner, E.F. (2003). AP-1: a double-edged sword in tumorigenesis. Nat. Rev. Cancer *3*, 859–868. https://doi.org/10.1038/nrc1209.

53. Shaulian, E., and Karin, M. (2002). AP-1 as a regulator of cell life and death. Nat. Cell Biol. *4*, E131–E136. https://doi.org/10.1038/ncb0502-e131.

54. Fujii, M., Iwai, K., Oie, M., Fukushi, M., Yamamoto, N., Kannagi, M., and Mori, N. (2000). Activation of oncogenic transcription factor AP-1 in T cells infected with human T cell leukemia virus type 1. AIDS Res. Hum. Retroviruses *16*, 1603–1606. https://doi.org/10.1089/08892220050193029.

55. Li, L., and Davie, J.R. (2010). The role of Sp1 and Sp3 in normal and cancer cell biology. Ann. Anat. *192*, 275–283. https://doi.org/10.1016/j.aanat.2010.07.010.

56. Wessner, R., Yao, J., and Wigdahl, B. (1997). Sp family members preferentially interact with the promoter proximal repeat within the HTLV-I enhancer. Leukemia *11* (*Suppl 3*), 10–13.

57. Liang, J., Cui, Z., Wu, C., Yu, Y., Tian, R., Xie, H., Jin, Z., Fan, W., Xie, W., Huang, Z., et al. (2021). DeepEBV: a deep learning model to predict Epstein–Barr virus (EBV) integration sites. Bioinformatics *37*, 3405–3411. https://doi.org/10.1093/bioinformatics/btab388.

58. Tian, R., Zhou, P., Li, M., Tan, J., Cui, Z., Xu, W., Wei, J., Zhu, J., Jin, Z., Cao, C., et al. (2021). DeepHPV: a deep learning model to predict human papillomavirus integration sites. Brief. Bioinform. *22*, bbaa242. https://doi.org/10.1093/bib/bbaa242.

59. Wu, C., Guo, X., Li, M., Shen, J., Fu, X., Xie, Q., Hou, Z., Zhai, M., Qiu, X., Cui, Z., et al. (2021). DeepHBV: a deep learning model to predict hepatitis B virus (HBV) integration sites. BMC Ecol. Evol. *21*, 138. https://doi.org/10.1186/s12862-021-01869-8.

60. Li, L., He, S., Sun, J.M., and Davie, J.R. (2004). Gene regulation by Sp1 and Sp3. Biochem. Cell. Biol. *82*, 460–471. https://doi.org/10.1139/o04-045.

61. Hagen, G., Müller, S., Beato, M., and Suske, G. (1992). Cloning by recognition site screening of two novel GT box binding proteins: a family of Sp1 related genes. Nucleic Acids Res. *20*, 5519–5525. https://doi.org/10.1093/nar/20.21.5519.

62. Liang, H., Xiao, G., Yin, H., Hippenmeyer, S., Horowitz, J.M., and Ghashghaei, H.T. (2013). Neural development is dependent on the function of specificity protein 2 in cell cycle progression. Development *140*, 552–561. https://doi.org/10.1242/dev.085621.

63. Johnson, C.A., and Ghashghaei, H.T. (2020). Sp2 regulates late neurogenic but not early expansive divisions of neural stem cells underlying population growth in the mouse cortex. Development *147*, dev186056. https://doi.org/10.1242/dev.186056.

64. Ramos, B., Gaudillière, B., Bonni, A., and Gill, G. (2007). Transcription factor Sp4 regulates dendritic patterning during cerebellar maturation. Proc. Natl. Acad. Sci. USA *104*, 9882–9887. https://doi.org/10.1073/pnas.0701946104.

65. Pinacho, R., Villalmanzo, N., Lalonde, J., Haro, J.M., Meana, J.J., Gill, G., and Ramos, B. (2011). The transcription factor SP4 is reduced in postmortem cerebellum of bipolar disorder subjects: control by depolarization and lithium. Bipolar Disord. *13*, 474–485. https://doi.org/10.1111/j.1399-5618.2011.00941.x.

66. Ratajewski, M., Walczak-Drzewiecka, A., Gorzkiewicz, M., Sałkowska, A., and Dastych, J. (2016). Expression of human gene coding RORγT receptor depends on the Sp2 transcription factor. J. Leukoc. Biol. *100*, 1213–1223. https://doi.org/10.1189/jlb.6A0515-212RR.

67. Leal, F.E., Ndhlovu, L.C., Hasenkrug, A.M., Bruno, F.R., Carvalho, K.I., Wynn-Williams, H., Neto, W.K., Sanabani, S.S., Segurado, A.C., Nixon, D.F., and Kallas, E.G. (2013). Expansion in CD39+ CD4+ immunoregulatory T cells and rarity of Th17 cells in HTLV-1 infected patients is associated with neurological complications. PLoS Negl. Trop. Dis. *7*, e2028. https://doi.org/10.1371/journal.pntd.0002028.

68. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

69. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

70. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

71. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

72. O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., and others. (2019). KerasTuner (github.Com/keras-Team/keras-Tuner).

73. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838. https://doi.org/10.1038/nbt.3300.

74. Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999. https://doi.org/10.1101/gr.200535.115.

75. Budach, S., and Marsico, A. (2018). pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. Bioinformatics *34*, 3035–3037. https://doi.org/10.1093/bioinformatics/bty222.